ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 3 (2), 2023

https://doi.org/10.23939/ictee2023.02.020

МЕТОДИКА КРИТЕРІЇВ СУМ У ЗАДАЧАХ ТЕСТУВАННЯ НЕЗАЛЕЖНОСТІ ПОСЛІДОВНОСТЕЙ ВИПАДКОВИХ ЧИСЕЛ

М. Одегов [ORCID:0000-0001-5526-2487], Ю. Бабіч [ORCID: 0000-0002-7888-7591], Д. Багачук [ORCID: 0000-0002-7888-7591], М. Кочеткова, Я. Петрович

Державний університет інтелектуальних технологій і зв'язку, вул. Кузнечна, 1, Одеса, 65000, Україна

Відповідальний за рукопис: Микола Одегов (onick_64@ukr.net)

(Подано 28 Липня 2023)

Генератори випадкових та псевдовипадкових чисел (ГВЧ) спочатку використовували для задач числового інтегрування (метод Монте-Карло). Нині основними сферами застосування ГВЧ є імітаційне моделювання та криптографія. Для першої сфери характерне використання ГВЧ, основаних на використанні комп'ютерних алгоритмів і програм. У статті розглянуто методику тестування незалежності послідовностей випадкових чисел (ПВЧ). Методика основана на властивостях сум незалежних випадкових величин. Алгоритми за цією методикою відповідають умові великої швидкості. Проаналізовано не лише моментні статистики типу коефіцієнтів кореляції, а й властивості емпіричних функцій розподілення сум ПВЧ. У статті аналіз обмежений лише випадком рівномірно розподілених ПВЧ. Виконані розрахунки доводять високу селективну ефективність запропонованих критеріїв, які дають змогу впевнено розрізняти залежні та незалежні ПВЧ. Завдяки швидкості запропоновані алгоритми та критерії можна використовувати для тестування ПВЧ надвеликої довжини (у задачах типу Big Data).

Ключові слова: псевдовипадкові числа; незалежність; суми незалежних послідовностей; непараметричні критерії; рівномірне розподілення.
УДК 004.94

1. Вступ

Сфера застосування послідовностей випадкових чисел (ПВЧ) постійно розширюється. Сьогодні лише у найзагальнішому вигляді області застосування ПВЧ визначаються як [1]: імітаційне моделювання, криптографія, генерація захищених ключів та ігрові задачі. Залежно від типу задачі використовують генератори випадкових та псевдовипадкових чисел (відповідно, ГВЧ та ГПВЧ). ГВЧ ефективно застосовують у криптографічних задачах, де зазвичай використовують ті чи інші фізичні явища. Цікаво, що навіть після відповідної обробки енцефалограм людини можна отримати ПВЧ [2]. Кількість наукових робіт, що стосуються реалізації ГВЧ та ГПВЧ, конкретних задач із використанням ПВЧ, а також тестування відповідних рішень постійно зростає. Є навіть принципова можливість отримати ПВЧ з використанням хмарних технологій [3].

Методика, розглянута у цій роботі, призначена передусім для тестування ПВЧ, які застосовують у задачах імітаційного моделювання [4]. На наш погляд, для цього класу задач найефективнішим вважають використання саме ГПВЧ, оскільки отримані ПВЧ можна відтворити у повторних числових експериментах. ГПВЧ можуть бути реалізовані як апаратними [5], так і

програмними [4] засобами. В останньому випадку застосовують різні рекурентні алгоритми [6, 7], але найпоширенішими залишаються алгоритми, основані на конгруентних методах першого та вищих порядків [4, 8].

Серед задач імітаційного моделювання типовою задачею є синтез моделей випадкових процесів із заданими характеристиками, наприклад, моделювання "кольорового" (не "білого") шуму [9]. Втім, на наш погляд, задача синтезу ПВЧ із заданою кореляційною функцією є менш складною, ніж задача синтезу некорельованих послідовностей: з других легко отримати перші за допомогою елементарних перетворень. Ще складнішою є задача синтезу незалежних ПВЧ.

У доволі простих тестах, окрім лінійних залежностей (кореляцій), аналізують і моментні статистики ПВЧ вищих порядків [10, 11]. Найчастіше тестування ПВЧ розуміють згідно із парадигмою "достатньої випадковості" [12], наприклад, за допомогою системи тестів американського стандарту NIST 800-22, 1a [13]. Цей стандарт містить 15 тестів, серед яких тест відсутності (малості) прихованих періодичностей. У роботі [14] зазначено, що ця система містить залежні тести, а також недостатньо обґрунтована теоретично. У цій роботі кількість тестів зменшена до шести і подано логічне і послідовне обґрунтування цієї системи (SixTestsSuite). Далі розглядатимемо лише один тест – незалежності ПВЧ.

Метою цієї роботи є обґрунтування методики тестування незалежності ПВЧ, основаної на властивостях сум незалежних випадкових величин та на використанні непараметричних критеріїв типу Колмогорова, Смірнова та Крамера – Мізеса – Смірнова (критерій ω^2).

2. Теоретичні основи запропонованої методики

У роботі досліджуватимемо саме незалежність ПВЧ. Оскільки ПВЧ у задачах імітаційного моделювання відтворюють (з певною точністю) теоретичні розподілення, то і теоретичний аналіз спочатку виконаємо для моделей теорії ймовірностей [15, 16]. За класичними визначеннями випадкові величини (ВВ) є незалежними тоді й тільки тоді, коли:

$$P(X,Y) = P(X) * P(Y), \tag{1}$$

де P ймовірності конкретних значень ВВ X та Y, або за еквівалентним визначенням:

$$G(x,y) = G_1(x) * G_2(y), \tag{2}$$

де G – функції розподілення (ФР) ймовірностей відповідних ВВ.

Зрозуміло, що для ПВЧ, які моделюють дискретні розподілення із порівняно невеликою кількістю дискрет, визначення (1) може бути продуктивним. Втім, для безперервних ВВ здебільшого воно надто складне для безпосереднього використання. Справді, окремі значення ВВ безперервних значень X та Y (у ПВЧ) іноді взагалі не збігаюьбся, а в результаті можна отримати матрицю значень $P(X,Y) = P(X) * P(Y)$, в якій кожний елемент просто дорівнює нулю. Тому треба використовувати оцінки частот потрапляння елементів ПВЧ у якісь кластери, наприклад, квантильні. А тут виникають інші проблеми:

– ширину кластерів дослідник задає практично навмання, що у певних випадках обмежує застосування автоматичних алгоритмів;

– кількість операцій для розв'язання задачі матиме порядок N^2, де N – типова довжина (кількість значень) ПВЧ, які моделюють ВВ X та Y.

Типові значення довжини неповторюваних ПВЧ, які генерує ГПВЧ, дорівнює 2^{32} та 2^{64} залежно від бібліотеки стандартних функцій тієї чи іншої мови програмування. Втім, значення цих порядків можуть бути і значно більшими. У задачах імітаційного моделювання ці ПВЧ використовують повністю або частково. Відповідно, обсяг тестувань може бути значним (типова задача класу BigData). Іноді під час розв'язання задач такого типу доводиться жертвувати якістю рішень замість можливості отримати ці рішення у прийнятний час [17].

Розглянемо лише випадок суцільно безперервних ВВ, для якого умови незалежності (1, 2) можна формалізувати в еквівалентному вигляді:

$$g(x,y) = g_1(x) * g_2(y), \tag{3}$$

де g – щільності розподілення відповідних ВВ.

Для цього випадку із умови незалежності (3) випливає відома рівність для суми $z = x + y$ незалежних ВВ [16]:

$$g(z) = \int_{-\infty}^{\infty} g_1(x) * g_2(z-x)dx, \tag{4}$$

тобто щільність суми виражається як згортка щільностей ВВ X та $Y\left(g(z) = g_1 * g_2\right)$. Важливо, що аналіз незалежності у цьому випадку вдвічі скорочує розмірність простору рішень: замість дослідження властивостей двох змінних (1)–(3) досліджуються функції лише однієї змінної. Виграш у кількості операцій очевидний.

Подальші дослідження ґрунтуються саме на залежності (4). На жаль, не вдалось знайти зворотних теорем: за яких умов із залежності (4) випливає залежність (3). Проте не знайшлось і прикладів, коли зворотна імплікація неправильна. Моделювання на різних прикладах (за умови $g_1 \equiv g_2$) показало, що твердження (3) та (4) тотожні. Втім, заради математичної строгості надалі вважатимемо, що співвідношення (4) є лише необхідною умовою незалежності ВВ, але у загальному випадку недостатньою. Переваги та недоліки пропонованої методики, основаної на залежності (4), розглянемо на конкретних прикладах.

3. Алгоритм аналізу незалежності рівномірно розподілених ПВЧ (РР ПВЧ)

Для багатьох методів генерування ПВЧ датчики РР ПВЧ є базовими [4], тому саме цей випадок заслуговує на увагу. Надалі ФР рівномірно розподілення ВВ на інтервалі [0, 1] будемо позначати U (uniform), а щільність розподілення, відповідно u. Тоді умову (4) для однакових незалежних рівномірно розподілених ВВ можна записати у вигляді: $u^*(x+y) = u^*(z) = u(x) * u(y)$. Для цього випадку функція згортки має тривіальний вигляд:

$$u^*(z) = \begin{cases} 0, z < 0 \\ z, 0 \le z < 1 \\ 2 - z, 1 \le z < 2 \\ 0, z > 2 \end{cases}, \tag{5}$$

а графіки щільності u та її автозгортк" $g(z)$ подано на рис. 1.

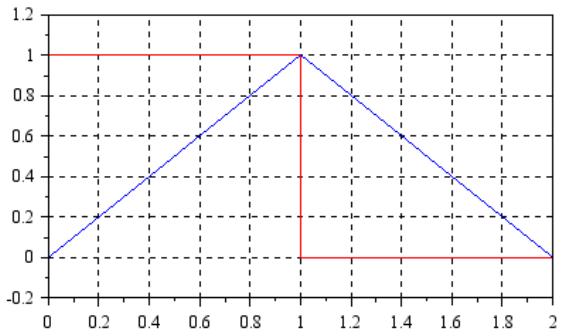

Рис. 1. Рівномірна щільність розподілення на інтервалі [0, 1] та її "автозгортка"

Безпосереднє використання моделей типу (5) у вигляді щільностей розподілення сум має певний недолік: емпіричним аналогом щільності є гістограма. Вище зазначено, що у цьому випадку результат залежить від інтуїції людини-дослідника. Тому надалі оперуватимемо поняттям функцій розподілення ВВ та, відповідно, емпіричних функцій розподілення (ЕФР). Останні визначають відомим способом:

$$G_N(x) = \frac{1}{N}\sum_{n=1}^{N} H(x - X_N), \tag{6}$$

де H – одинична функція Хевісайда; X_N – вибіркові значення ВВ.

Інтегруванням залежностей (5) отримаємо ФР суми незалежних ВВ, рівномірно розподілених на інтервалі [0, 1]:

$$U^*(z) = \begin{cases} 0, z < 0 \\ 0.5 * z^2, 0 \le z < 1 \\ -1 + 2z - 0.5z^2, 1 \le z < 2 \\ 0, z > 2 \end{cases}, \tag{7}$$

графік якої показано на рис. 2.

Тоді у загальному вигляді пропонований алгоритм складатиметься з таких операцій:

1. Отримуємо масив РР ПВЧ $X_1, X_2, ...X_N$ загальною довжиною $2{\cdot}N$.

2. Визначаємо масив парних сум $Z_1 = X_1 + X_{N+1}$, $Z_2 = X_2 + X_{N+2}$, і т.д. загальною довжиною N.

3. Визначаємо ЕФР $G_N(z)$ за правилом (6), порівнюючи її з теоретичною функцією розподілення (7).

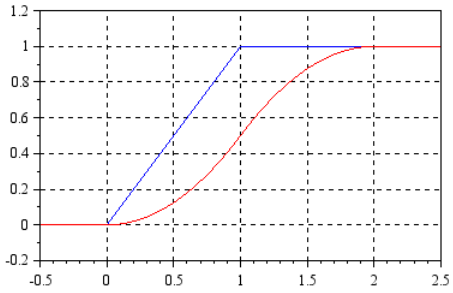

Рис. 2. ФР рівномірного розподілення на інтервалі [0, 1] та ФР суми незалежних ВВ

Приклади ЕФР сум РР РВЧ для довжини послідовностей $N = 20$ та $N = 50$ подано на рис. 3.

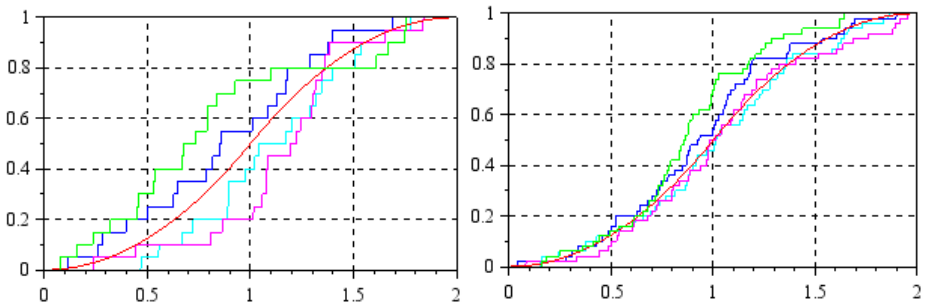

Рис. 3. Приклади реалізації ЕФР сум для N = 20 (а) та N = 50 (б)

Але принципове і вже зовсім нетривіальне запитання: які показники розбіжності або подібності ЕФР $G_N(z)$ та теоретичної функції $U^*(z)$ вибрати? Розглянемо це питання докладно.

4. Обґрунтування вибору показників та критеріїв незалежності РР ПВЧ

Для задач аналізованого класу доцільно використовувати непараметричні критерії, основані на статистиках відстаней між теоретичною ФР та ЕФР. У цій роботі розглянемо лише три непараметричні статистики: Колмогорова, "однобічну вибіркову" статистику (ОВС) Смірнова та статистику ω^2 (Крамера – Мізеса – Смірнова).

Статистика Колмогорова [15] ґрунтується на зваженій метриці Чебишева:

$$R_K(N) = \sqrt{N}\sup_x\left|G_N(x) - G(x)\right|. \tag{8}$$

Відомо [15], що за $N \to \infty$ ймовірність нерівності $R_K(N) < \lambda$ рівномірно сходиться до функції:

$$K(\lambda) = \sum_{k=-\infty}^{k=\infty}(-1)^k \exp\left(-2k^2\lambda^2\right). \tag{9}$$

Перевагою статистики (8) для використання у експрес-методах є саме можливість дуже швидко отримати її значення: потрібно лише визначити масив різниць на певній сітці моделювання, обчислити модулі цих різниць і знайти найбільшу з-поміж них.

Певні недоліки полягають у тому, що визначення конкретних значень за функцією (9) може тривати довго. Також зауважимо, що функція (9) більш-менш придатна для розв'язання задач на великих вибірках (відповідно до способу її визначення), але для малих вибірок вигляд функції ще складніший.

ОВС Смірнова виражається залежністю[18], яку подамо в еквівалентній формі:

$$R_S(N) = \sqrt{N}\sup_x\left(G_N(x) - G(x)\right). \tag{10}$$

Статистика (10) ще більш придатна для оброблення даних типу BigData (де доводиться рахувати кожну мікросекунду в циклічних операціях): порівняно з виразом (8), не потрібно навіть змінювати значення знакових бітів. Втім, зауважимо, що подібність визначень (8) та (10) далека від математичної суті: у другому випадку аналізують лише перевищення значень ЕФР над теоретичною ФР (звідки і термін "однобічна").

Зрозуміло, що внаслідок властивостей ФР завжди $R_S(N) \ge 0$ на інтервалі $\left[-\infty, \infty\right]$. А що робити, якщо у конкретному обчислювальному експерименті отримано значення: $R_S(N) = 0$? Теоретично можливі два випадки: ЕФР ідеально збігається з теоретичною ФР та альтернативний випадок, коли ЕФР практично всюди значно менша, ніж теоретична ФР. Відокремлення цих двох випадків розглянемо нижче, а поки зазначимо, що ймовірність нерівності $R_S(N) < \lambda$, якщо $N \to \infty$, рівномірно сходиться до функції [18]:

$$S(\lambda) = P\left(R_S(N) < \lambda = 1 - \exp\left(-2*\lambda^2\right)\right),\ \lambda \ge 0 \tag{11}$$

звідки щільність розподілення (після додаткового визначення $S(\lambda) = 0,\ \lambda < 0$)

$$S(\lambda) = \frac{dS(\lambda)}{d\lambda} = \begin{cases} 4\lambda\exp\left(-2\lambda^2\right), \lambda \ge 0 \\ 0, \lambda < 0 \end{cases}. \tag{12}$$

Графіки функцій (11) та (12) показано на рис. 4.

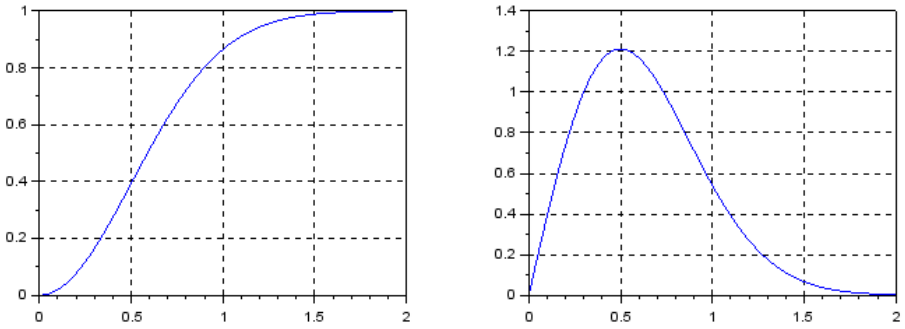

Рис. 4. Функція розподілення (11) та щільність розподілення (12)

Простота, так би мовити, "математична елегантність" виразів (11) та (12) дає змогу елементарно визначити математичне очікування (M_S), дисперсію (D_S) та стандартне відхилення (σ_S) для цих функцій:

$$M_S = \frac{1}{2}\sqrt{\frac{\pi}{4}} \approx 0.626657;\ D_S = \frac{1}{2}\left(1 - \frac{\pi}{4}\right) \approx 0.107301;\ \sigma_S = \sqrt{D_S} \approx 0.327568, \qquad (13)$$

а також доволі просто виразити квантилі порядку $\frac{q}{Q}$ на сітці Q квантилів, розв'язуючи рівняння, яке випливає з виразу (11):

$$\lambda\left(\frac{q}{Q}\right) = \sqrt{-\frac{1}{2}\ln\left(1 - \frac{q}{Q}\right)}, \qquad (14)$$

яке, наприклад, для медіани має простий вигляд: $\lambda(0.5) = \sqrt{-0.5\ln(0.5)} \approx 0.588705$.

Квантилі порядку $\frac{q}{10}$ (децилі) наведено у табл. 1, де також подано критичні значення для вибірки об'єму $N = 100$.

Таблиця 1

Квантилі функції розподілення статистики Смірнова (11)

Значення $\frac{q}{Q}$	$S\left(\frac{q}{Q}\right)$	$S\left(\frac{q}{Q}\right)/10$	Значення $\frac{q}{Q}$	$S\left(\frac{q}{Q}\right)$	$S\left(\frac{q}{Q}\right)/10$
0,1	0,229522	0,022952	0,6	0,676864	0,067686
0,2	0,334024	0,033402	0,7	0,775878	0,077588
0,3	0,422300	0,042230	0,8	0,897061	0,089706
0,4	0,505384	0,050538	0,9	1,072983	0,107298
0,5	0,588705	0,058875	0,994372	1,609362	0,160936

У табл. 1 є окреме значення "квантиля" 0,994372. Воно подано замість значення $\frac{q}{Q} = 1$, оскільки для $\frac{q}{Q} \to 1$ $S\left(\frac{q}{Q}\right) \to \infty$. Втім, це значення у такому контексті не є випадковим. Для експрес-аналізу часто доцільно використовувати нерівність Чебишева у вигляді правила "трьох сигм". З урахуванням асиметрії щільності розподілення (рис. 4) та "однобічності" статистики (10) критичні значення $R_S(N)$ за цим правилом можна виразити:

$$R_S(N) = \frac{1}{\sqrt{N}}\left(M_S + 3\sigma_S\right) \approx 1.609362\frac{1}{\sqrt{N}} \approx \frac{1.6}{\sqrt{N}}, \quad (15)$$

що і відображено у табл. 1.

Отже, у швидких алгоритмах аналізу незалежності РР ПВЧ можна використовувати два критичні значення: квантиль 90 % (приблизно $1/\sqrt{N}$) та "оцінку Чебишева", яка у такому разі дає ймовірність виходу за критичне значення не більше ніж шість разів на 1000 випадків.

Враховуючи отримані числові значення, можна сформулювати послідовність критеріальних правил:

– критерій 0,1: якщо $1/\sqrt{N} < R_S(N)$ то гіпотеза незалежності ПВЧ відкидається з ймовірністю помилки рішення приблизно 0,1;

– критерій 0,006: якщо $1.6/\sqrt{N} < R_S(N)$ то гіпотеза незалежності ПВЧ відкидається з ймовірністю помилки рішення приблизно 0,006 (шість тисячних);

– випадок 0: якщо $R_S(N) = 0$ то робимо висновок, що "однобічна статистика" (10) не дає адекватного результату і потрібно використовувати якісь інші критерії.

Один із таких можливих "інших" критеріїв можна побудувати на статистиці ω^2 [19], яка визначається у метриці простору L_2:

$$R_\omega(N) = \int_{-\infty}^{\infty}\left[G_N(x) - G(x)\right]^2 dG(x). \quad (16)$$

Функція розподілення значень $R_\omega(N)$ залежить від об'єму вибірки N, тому заздалегідь визначити якісь квантилі складно. Якщо скористатись граничними значеннями цієї функції за умови $N \to \infty$, то втрачається можливість коректного використання цієї статистики для малих вибірок. Втім, відомі значення математичного очікування та дисперсії цієї статистики для конкретного об'єму вибірки N [19], а саме:

$$M_\omega(N) = \frac{1}{6N\sqrt{N}};\ D_N = \frac{1}{\sqrt{N}} * \frac{4N-3}{180N^2}. \quad (17)$$

Тоді навіть для малих вибірок можна застосовувати критерії на основі нерівності Чебишева. Втім, у цьому випадку скористаємось узагальненим правилом "M сигм". Критичні значення для прийняття рішень про порушення умов незалежності з ймовірністю помилки приблизно $\frac{1}{M^2}$ можна приймати за спрощеною формулою (якщо $N \geq 10$):

$$K_\omega(M, N) = \frac{1}{\sqrt{N}}\left(\frac{1}{6N} + \frac{M}{3}\sqrt{\frac{1}{5N}}\right). \quad (18)$$

Використовуючи формулу (18), обмежимось лише двома правилами: "трьох сигм" та "шести сигм". Зауважимо, що навіть за порівняно невеликих об'ємів вибірок N значення $K_\omega(6, N)$ приблизно вдвічі більше, ніж значення $K_\omega(3, N)$. Тоді критеріальні правила можна сформулювати так:

– критерій 3σ: якщо $K_\omega(N) > K_\omega(3, N)$, то гіпотеза незалежності ПВЧ відкидається з ймовірністю помилки рішення приблизно 0,1;

– критерій 6σ: якщо $K_{\omega}(N) > K_{\omega}(6, N)$, то гіпотеза незалежності ПВЧ відкидається з ймовірністю помилки рішення приблизно 0,03.

Звернемо увагу, що у формулюванні критеріальних правил ми дуже обережно використовували поняття ймовірності, а також користувались принципом, так би мовити, "презумпції залежності". На наше переконання, поняття "довірчої ймовірності" не завжди застосовне до конкретної статистичної задачі: можна знайти такий рівень цього показника, що будь-яке емпіричне розподілення можна буде вважати, наприклад, нормальним. Тому в практичних дослідженнях принципово важливим є контроль отриманих висновків на додаткових (незалежних) масивах експериментальних даних. Скільки таких контрольних вибірок треба отримати? Чим більше, тим краще, але глобальний мінімум – хоча б одну. На щастя, ГВЧ дають практично необмежену можливість контролю одержаних рішень.

5. Перевірка незалежності РР ПВЧ у середовищі Scilab за допомогою критеріїв сум

Всі графіки і розрахунки вище отримано у середовищі Scilab. Надалі досліджуватимемо характеристики залежності РР ПВЧ саме у цьому середовищі. Для цього будемо використовувати стандартний ГПВЧ Scilab із параметром налаштування "uniform".

Для обчислювальних експериментів генеруються дві РР ПВЧ: $X_n, n = \overline{1, N}$ та $Y_n, n = \overline{1, N}$, з яких для подальшого аналізу формується масив сум: $Z_n = X_n + Y_n, n = \overline{1, N}$. Розглядають три випадки:

– "умовна незалежність": ПВЧ X та Y формуються як послідовні масиви випадкових чисел, які не перетинаються;

– "повна залежність": $X_n \equiv Y_n \Rightarrow Z_n = 2X_n, n = \overline{1, N}$;

– "напівзалежність": половина ПВЧ Y заповнюється значеннями з ПВЧ X, а ще одна половина – незалежною ПВЧ V завдовжки $N/2$, тобто $X_{2n-1} = 2X_{2n-1}, Z_{2n} = X_{2n} + V_n, n = \overline{1, N/2}$.

Графіки сум для цих трьох випадків подано на рис. 5.

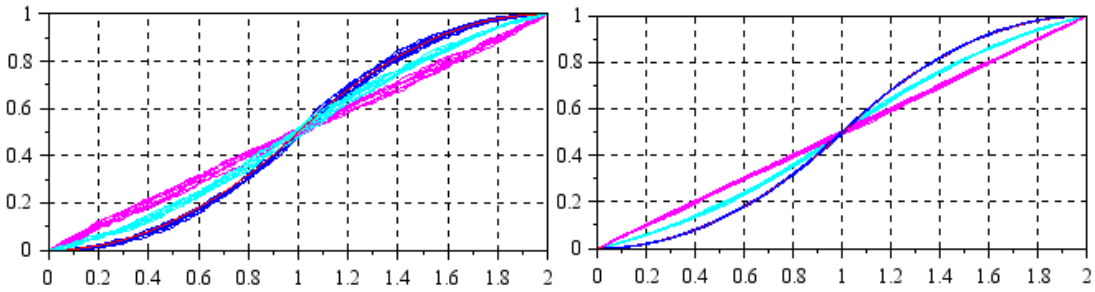

Рис. 5. Графіки сум РР ПВЧ для трьох випадків залежності та для довжини ПВЧ N = 1000 та N = 10000

Зауважимо, що для випадків "повної залежності" та "напівзалежності" корисне таке математичне положення.

Лема. Нехай дано дві некорельовані ВВ X та Y $(r(X, Y) = 0)$ із однаковою ФР. Сконструюємо третю ВВ як зважену суму: $V = \alpha X + (1 - \alpha)Y, 0 \le \alpha \le 1$. Тоді $r(V, X) = \alpha$.

Доказ цієї леми не наводимо, зважаючи на її елементарність.

Результати обчислювального експерименту для випадку "умовної незалежності" та порівняно невеликої довжини ПВЧ ($N = 100$) наведено у табл. 2. У цій таблиці, як і в усіх наступних таблицях,

подано результати десяти незалежних експериментів (тестів). Також у таблицях довідково наведено значення коефіцієнта кореляції $r(V,X)$ базових послідовностей X та Y.

У табл. 3 подано результати аналогічного обчислювального експерименту, але для порівняно великої довжини ПВЧ *N*= 10000.

Таблиця 2

Результати обчислювального експерименту для випадку "умовної незалежності" (*N* = 100)

№ тесту	ОВС Смірнова			Статистика ω^2			$r(X,Y)$
	Крит. 0,1	Крит. 0,006	$R_s(N)$	$K_\omega(3,N)$	$K_\omega(6,N)$	$R_\omega(N)$	
1	0,1	0,16	0,0678365	0,0021667	0,0043333	0,0006823	0,039205
2	0,1	0,16	0,0910911	0,0021667	0,0043333	0,0027850	–0,136566
3	0,1	0,16	0,0217520	0,0021667	0,0043333	0,0031126	0,107362
4	0,1	0,16	0,0267011	0,0021667	0,0043333	0,0007893	0,064931
5	0,1	0,16	0,0377856	0,0021667	0,0043333	0,0024807	–0,008505
6	0,1	0,16	0,0303478	0,0021667	0,0043333	0,0004274	0,069523
7	0,1	0,16	0,0341514	0,0021667	0,0043333	0,0014580	–0,144585
8	0,1	0,16	0,0348938	0,0021667	0,0043333	0,0005420	0,011181
9	0,1	0,16	0,0508711	0,0021667	0,0043333	0,0005627	0,105078
10	0,1	0,16	0,1073036	0,0021667	0,0043333	0,0029783	–0,075195
Ср. знач.	0,1	0,16	0,0502734	0,0021667	0,0043333	0,0015818	0,003242

Таблиця 3

Результати обчислювального експерименту для випадку "умовної незалежності" (N= 10000)

№ тесту	ОВС Смірнова			Статистика ω^2			$r(X,Y)$
	Крит. 0,1	Крит. 0,006	$R_s(N)$	$K_\omega(3,N)$	$K_\omega(6,N)$	$R_\omega(N)$	
1	0,01	0,016	0,0043159	0,0000202	0,0000403	0,0000058	-0,002522
2	0,01	0,016	0,0077291	0,0000202	0,0000403	0,0000139	0,000422
3	0,01	0,016	0,0077928	0,0000202	0,0000403	**0,0000242**	0,002253
4	0,01	0,016	0,0066080	0,0000202	0,0000403	0,0000116	0,004661
5	0,01	0,016	**0,0105816**	0,0000202	0,0000403	0,0000179	-0,003972
6	0,01	0,016	0,0082445	0,0000202	0,0000403	0,0000134	-0,000291
7	0,01	0,016	0,0055198	0,0000202	0,0000403	0,0000135	0,001396
8	0,01	0,016	0,0049011	0,0000202	0,0000403	0,0000072	-0,005238
9	0,01	0,016	0,0064132	0,0000202	0,0000403	0,0000118	-0,001651
10	0,01	0,016	0,0057517	0,0000202	0,0000403	0,0000071	-0,011419
Ср. знач.	0,01	0,016	0,0067858	0,0000202	0,0000403	0,0000126	-0,001636

Тут і далі виділено клітини, де значення показників є проміжними між двома критеріями: $1/\sqrt{N} < R_S(N) < 1.6/\sqrt{N}$ для ОВС Смірнова та $K_\omega(3,N) < R_\omega(N) < K_\omega(6,N)$ для статистики ω^2. Іншим кольором виділено протилежні значення залежно від того, що доводять: залежність або незалежність ПВЧ.

Аналогічні результати для випадку повної залежності наведено у табл. 4 та табл. 5. Вочевидь, у даному випадку коефіцієнт кореляції завжди: $r(X,Y)=1$, тобто згідно Леми, $\alpha=1$.

У табл. 6 та табл. 7 наведено результати для випадку "напівзалежності". Тут, згідно із лемою, $\alpha=0{,}5$ і очікуване значення коефіцієнта кореляції має становити приблизно $r(X,Y)\approx 0{,}5$.

Таблиця 4

Результати обчислювального експерименту для випадку "повної залежності" (N= 100)

№ тесту	ОВС Смірнова			Статистика ω^2			$r(X,Y)$
	Крит. 0,1	Крит. 0,006	$R_s(N)$	$K_\omega(3,N)$	$K_\omega(6,N)$	$R_\omega(N)$	
1	0,1	0,16	**0,1116782**	0,0021667	0,0043333	0,0243284	1,0
2	0,1	0,16	0,1856628	0,0021667	0,0043333	0,0230020	1,0
3	0,1	0,16	**0,1474957**	0,0021667	0,0043333	0,0146878	1,0
4	0,1	0,16	0,2149174	0,0021667	0,0043333	0,0252370	1,0
5	0,1	0,16	**0,1543380**	0,0021667	0,0043333	0,0256452	1,0
6	0,1	0,16	0,1758654	0,0021667	0,0043333	0,0245380	1,0
7	0,1	0,16	**0,1016281**	0,0021667	0,0043333	0,0098269	1,0
8	0,1	0,16	0,1707647	0,0021667	0,0043333	0,0347024	1,0
9	0,1	0,16	0,2135089	0,0021667	0,0043333	0,0370261	1,0
10	0,1	0,16	**0,1379820**	0,0021667	0,0043333	0,0127843	1,0
Ср. знач.	0,1	0,16	0,1613841	0,0021667	0,0043333	0,0231778	1,0

Таблиця 5

Результати обчислювального експерименту для випадку "повної залежності" (N = 10000)

№ тесту	ОВС Смірнова			Статистика ω^2			$r(X,Y)$
	Крит. 0,1	Крит. 0,006	$R_s(N)$	$K_\omega(3,N)$	$K_\omega(6,N)$	$R_\omega(N)$	
1	0,01	0,016	0,1270820	0,0000202	0,0000403	0,0175547	1,0
2	0,01	0,016	0,1256242	0,0000202	0,0000403	0,0175356	1,0
3	0,01	0,016	0,1353969	0,0000202	0,0000403	0,0171425	1,0
4	0,01	0,016	0,1330270	0,0000202	0,0000403	0,0168206	1,0
5	0,01	0,016	0,1217417	0,0000202	0,0000403	0,0160211	1,0
6	0,01	0,016	0,1280348	0,0000202	0,0000403	0,0177882	1,0
7	0,01	0,016	0,1247589	0,0000202	0,0000403	0,0164111	1,0
8	0,01	0,016	0,1311875	0,0000202	0,0000403	0,0170850	1,0
9	0,01	0,016	0,1254439	0,0000202	0,0000403	0,0158164	1,0
10	0,01	0,016	0,1243558	0,0000202	0,0000403	0,0166027	1,0
Ср. знач.	0,01	0,016	0,1276653	0,0000202	0,0000403	0,0168778	1,0

Таблиця 6

Результати обчислювального експерименту для випадку "напівзалежності" (N = 100)

№ тесту	ОВС Смірнова			Статистика ω^2			$r(X,Y)$
	Крит. 0,1	Крит. 0,006	$R_s(N)$	$K_\omega(3,N)$	$K_\omega(6,N)$	$R_\omega(N)$	
1	0,1	0,16	**0,1166655**	0,0021667	0,0043333	0,0077644	0,509980
2	0,1	0,16	**0,1052469**	0,0021667	0,0043333	0,0129689	0,649185
3	0,1	0,16	**0,0826922**	0,0021667	0,0043333	0,0043945	0,578532
4	0,1	0,16	**0,1562606**	0,0021667	0,0043333	0,0079536	0,513632
5	0,1	0,16	**0,1081355**	0,0021667	0,0043333	0,0056331	0,468169
6	0,1	0,16	**0,1068974**	0,0021667	0,0043333	0,0092427	0,576968
7	0,1	0,16	**0,0705134**	0,0021667	0,0043333	**0,0037395**	0,558580
8	0,1	0,16	**0,1404395**	0,0021667	0,0043333	0,0093844	0,564223
9	0,1	0,16	**0,1033443**	0,0021667	0,0043333	0,0065026	0,586933
10	0,1	0,16	**0,1577742**	0,0021667	0,0043333	0,0108485	0,568619
Ср. знач.	0,1	0,16	**0,1147970**	0,0021667	0,0043333	0,0078432	0,557482

Таблиця 7

Результати обчислювального експерименту для випадку "напівзалежності" (N= 10000)

№ тесту	ОВС Смірнова			Статистика ω^2			$r(X,Y)$
	Крит. 0,1	Крит. 0,006	$R_S(N)$	$K_\omega(3,N)$	$K_\omega(6,N)$	$R_\omega(N)$	
1	0,01	0,016	0,0708602	0,0000202	0,0000403	0,0040883	0,500519
2	0,01	0,016	0,0641486	0,0000202	0,0000403	0,0041190	0,497059
3	0,01	0,016	0,0643166	0,0000202	0,0000403	0,0042779	0,507732
4	0,01	0,016	0,0660466	0,0000202	0,0000403	0,0040336	0,501539
5	0,01	0,016	0,0616841	0,0000202	0,0000403	0,0040181	0,500895
6	0,01	0,016	0,0641358	0,0000202	0,0000403	0,0041727	0,497858
7	0,01	0,016	0,0651654	0,0000202	0,0000403	0,0041838	0,501629
8	0,01	0,016	0,0609638	0,0000202	0,0000403	0,0037021	0,486865
9	0,01	0,016	0,0690960	0,0000202	0,0000403	0,0041380	0,493544
10	0,01	0,016	0,0537496	0,0000202	0,0000403	0,0038930	0,489873
Ср. знач.	0,01	0,016	0,0640167	0,0000202	0,0000403	0,0040627	0,497751

Проаналізуємо отримані результати.

Для випадку "умовної залежності" середні значення вказують, що за обома критеріями сум РР ПВЧ Scilab можна вважати незалежними, навіть для відносно невеликої довжини (N = 100, табл. 2). Втім, для малих вибірок отримуються і сумнівні значення. Так, у тестах 2, 3, 5 та 10 критерій ω^2 має проміжні значення $K_\omega(3,N) < R_\omega(N) < K_\omega(6,N)$. Тобто, гіпотеза незалежності відхиляється з ймовірністю помилкибільше 0,1, але не відкидається з ймовірністю помилки менше 0,006. Як бачимо, на 10 тестів таких випадків 4, тобто їх частота 0,4. Втім, за нерівністю Чебишева умовна ймовірність таких випадків має складати 0,1. В чому причина? Причина проста: нерівність Чебишева дає лише приблизні, граничні оцінки, які не завжди співпадають з квантилями ФР. Висновок: для вирішення задач тестування незалежності ПВЧ за допомогою критерія сум для статистики ω^2 треба орієнтуватись на верхнє значення $K_\omega(6,N)$.

У тій самій табл. 2 за критерієм ОВС Смірнова все гаразд: є рівно одне значення (тест №10), що виходить за межі $1/\sqrt{N} < R_S(N)$.

Результати для випадку "умовної незалежності" у табл. 3 (N = 10000) показують взагалі ідеальне співпадіння з теорією: пари ПВЧ вважаються незалежними у 9 випадках із 10, що відповідає рівням обох критеріїв 0,1. Також зауважимо, що у середньому обидва критерія сум впевнено вказують на незалежність ПВЧ.

Зовсім інші висновки потрібні у випадках явної залежності ПВЧ, яка може проявлятись також у відносно високому рівні їх лінійної залежності– кореляції. Тут бажано, щоб значення статистик $R_S(N)$ та $R_\omega(N)$ суттєво перевищували значення критичних границь незалежності.

Для випадку "повної залежності" у табл. 4 (N = 100) критерій ω^2 завжди дає адекватний результат навіть за верхнім значенням: $R_\omega(N) > K_\omega(6,N)$. Проте, критерій ОВС Смірнова у 50 % випадків дає лише виконання умови виходу за межу незалежності тільки для нижнього рівня. Втім, у середньому і даний критерій дає впевнений результат: ПВЧ залежні.

Для того ж випадку "повної залежності" згідно даних у табл. 5 для великої довжини ПВЧ (N = 10000) обидва критерія впевнено дають результат: ПВЧ залежні.

Найбільш цікавий у контексті даної роботи варіант "напівзалежності". Аналіз даних невеликої довжини ПВЧ у табл. 6 (N = 100) показує, що критерій ОВС Смірнова практично завжди відрізняє "залежність" і "незалежність" лише за верхнім рівнем критерію, але допускає і грубу

помилку (тести №3 та №7). Втім, у середньому, і цим критерієм можна користуватись з використанням верхнього рівня $1.6/\sqrt{N} < R_S(N)$.

Критерій ω^2 дає у даному випадку більш адекватні результати: лише в одному випадку (тест №7) виконується нерівність $K_\omega(3,N) < R_\omega(N) < K_\omega(6,N)$. В усіх інших випадках значення $R_\omega(N)$ перевищує верхній рівень, що у даному випадку і потрібно. Втім, як і у попередніх випадках, для великої довжини ПВЧ (табл. 7) обидва критерія дають однаково адекватні результати: ПВЧ залежні.

В цілому, аналіз результатів, що наведені вище, а також багатьох аналогічних показують, що для відносно малої довжини ПВЧ (порядку N = 100) більш доцільно користуватись критерієм ω^2, для середньої (N = 1000) і великої (N = 10000) довжини ПВЧ більш доцільним є критерій ОВС Смірнова. Останній дає найбільш швидкий алгоритм аналізу незалежності.

Висновки та рекомендації

Викладене дозволяє зробити наступні висновки.

1. Запропонований підхід тестування незалежності послідовностей випадкових чисел, заснований на критеріях сум та непараметричних статистиках має перевагу над методиками, що враховують лише моментні статистики, оскільки оперує саме розподіленнями.

2. Другою перевагою запропонованого підходу є надвисока швидкість обчислювальних алгоритмів, що дозволяє використовувати їх для вирішення задач обробки надвеликих послідовностей випадкових чисел (задачі класу BigData). Саме для даного класу задач дослідження у напрямку синтезу швидких алгоритмів є досить актуальною.

3. Теоретичним недоліком запропонованої методики є те, що досліджуються необхідні умови незалежності, але, у загальному випадку, не достатні.

4. Виконані обчислювальні експерименти показують, що навіть при відносно невеликих довжинах послідовностей випадкових чисел (порядку 100) використані критерії мають ознаки високої селективності: надійно відрізняють залежні та незалежні послідовності.

5. У даній роботі ми обмежились лише випадком рівномірно розподілених послідовностей. Втім, значна кількість методів генерації випадкових чисел з іншими розподіленнями генерується саме з рівномірного розподілення.

6. Генератор випадкових чисел Scilabвпевнено можна використовувати для моделювання випадкових процесів типу білого шуму, а також для моделювання різних явищ з контрольованими характеристиками залежності випадкових величин.

Список використаних джерел

[1] Herasymchuk O. I., Maksymovych V. M. (2003). "Generators of pseudo-random numbers, their application, classification, basic construction methods and quality assessment", Ukrainian Information Security Research Journal, Vol. 3, pp. 29–36. DOI: https://doi.org/10.18372/2410-7840.5.4270.

[2] Chen G. (2014). "Are electroencephalogram (EEG) signals pseudo-random number generators?", Journal of Computational and Applied Mathematics, Vol. 268, pp. 1–4, ISSN 0377-0427. DOI: https://doi.org/10.1016/j.cam.2014.02.028.

[3] Kumar V., Rayappan J., Amirtharajan R., Praveenkumar P. (2022). "Quantum true random number generation on IBM's cloud platform", Journal of King Saud University - Computer and Information Sciences, Vol. 34, Iss. 8, Part B, pp. 6453-6465, ISSN 1319-1578.DOI: https://doi.org/10.1016/j.jksuci.2022.01.015.

[4] Burtniak I.V. (2019). "Simulation modeling", Vasyl StefanykPrecarpathian National University, p. 97.

[5] Li Z., Li P., Mao Y., Halang W.A. (2005). Chaos-based Pseudo-Random Number Generators and Chip Implementation", IFAC Proceedings Volumes, Vol. 38, Issue 1,pp. 1090-1094, ISSN 1474-6670, ISBN 9783902661753. DOI: https://doi.org/10.3182/20050703-6-CZ-1902.00838.

[6] Akhshani A., Akhavan A., Mobaraki A., Lim S.-C., Hassan Z. (2014). "Pseudo random number generator based on quantum chaotic map", Communications in Nonlinear Science and Numerical Simulation, Vol. 19, Issue 1, pp. 101–111, ISSN 1007-5704. DOI: https://doi.org/10.1016/j.cnsns.2013.06.017.
[7] Sathya K., Sarveshwaran V., Subhika T., Devi M. (2022) "Security Analyses of Random Number Generation with Image Encryption Using Improved Chaotic Map", Procedia Computer Science, Vol. 215, pp. 432-441, ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.12.045.
[8] Pollard J.(1982). "Handbook of Computational Methods of Statistics", Finance and statistics, p. 344.
[9] Deza J.I., Ihshaish H. (2022). "qNoise: A generator of non-Gaussian colored noise", SoftwareX, Vol. 18, ISSN 2352-7110. DOI: https://doi.org/10.1016/j.softx.2022.101034.
[10] Luengo E.A., Cerna M., García Villalba L.J., Hernandez-Castro J. (2022). "A new approach to analyze the independence of statistical tests of randomness", Applied Mathematics and Computation, Vol. 426, ISSN 0096-3003. DOI: https://doi.org/10.1016/j.amc.2022.127116.
[11] Farmer J., Jacobs D. (2022)."MATLAB tool for probability density assessment and nonparametric estimation", SoftwareX, Vol. 18, ISSN 2352-7110. DOI: https://doi.org/10.1016/j.softx.2022.101017.
[12] Koivu A., Kakko J-P., Mäntyniemi S., Sairanen M. (2022). "Quality of randomness and node dropout regularization for fitting neural networks", Expert Systems with Applications, Vol. 207, ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.117938.
[13] Rukhin A., Soto J., Nechvatal J., Smid M., Barker E., Leigh S., Levenson M., Vangel M., Banks D., Heckert A., Dra J., Vo S. (2010). "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications: NIST Special Publication 800-22 Revision 1a", National Institute of Standards and Technology Gaithersburg, MD 20899-8930, p. 131.
[14] Kochana R., Kovalchuk L., Korchenko O.,Kuchynska N. (2021). "Statistical Tests Independence Verification Methods", Procedia Computer Science, Vol. 192, pp. 2678–2688, ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.09.038.
[15] Kartasjov M. V. (2008)."Probability, processes, statistics", Kyiv University, p. 494.
[16] Shyriaev A. N. (1980). "Probability", Science, p. 576.
[17] Odehov M. A., Hadzhyiev M. M., Bukata L. M., Hlazunova L. V., Kochetkova M. V. (2023). "Justification of fast classification algorithms on BIG DATA sets according to reliability and performance", Infocommunication and computer technologies. Iss. 1, pp. 148–160. DOI: https://doi.org /10.36994 / 2788-5518-2023-01-05-16.
[18] Orlov A. I. (2014). "Nonparametric goodness-of-fit tests by Kolmogorov, Smirnov, Omega-square and errors in their application", Science Magazine ofKubSAU, Iss. 97(03), p. 30.

SUM CRITERIA FOR THE TASK OF TESTING THE INDEPENDENCE OF RANDOM NUMBERS SEQUENCES

Nick Odegov, Yurii Babich, Denys Bahachuk, Maryna Kochetkova, Janna Petrovych

State university of intellectual technologies and communication, 1, Kuznechna str., Odesa, 65000, Ukraine

Random and pseudo-random number generators (RNGs) were initially used to solve numerical integration problems (the Monte Carlo method). Currently, the RNGs are used in cryptography and simulation modeling. The latter one typically uses RNGs based on computer algorithms and programs. This article presents a method aimed at testing the independence of random numbers sequences (RNSs). The method is based on the sums properties of independent random variables. Algorithms based on this method operate fast. Here not only the instant statistics including correlation coefficients are analyzed, but also the properties of empirical functions of RNSs distributed sums. In this article, the analysis is limited only to the case of uniformly distributed RNSs. The calculations performed prove the high selective efficiency of the proposed criteria, which allows to reliably distinguish between dependent and independent RNSs. Due to the high operation speed, the proposed algorithms and criteria can be used for testing very long RNSs (especially in Big Data tasks).

Key words: *pseudorandom numbers; independence; sums of independent sequences; non-parametric criteria; uniform distribution.*